Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 18:54:42
REFERENCE NO: 306

Author Names & Affiliations

- Jan Cheetham - University of Wisconsin-Madison

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Research Cyberinfrastructure, Office of the CIO

Title of Submission

The Role of Advanced Computing in Driving Research Innovation at the University of Wisconsin-Madison

Abstract (maximum ~200 words).

This document is a compilation of responses of 26 researchers (faculty, scientists, postdoctoral fellows, and graduate students) from the University of Wisconsin-Madison to the NSF's request for feedback on cyberinfrastructure needs to address big research questions and challenges in science and engineering over the next ten years. The disciplinary domains of these researchers span life sciences, plasma, colloidal, and astro physics, material sciences, economics, and clinical and bioinformatic sciences. Their responses demonstrate that computational resources are vital to the discoveries that are needed to advance research on new energy alternatives, materials, understandings of the interaction between genetics and the environment, and clinical therapies. While the availability of computational resources has provided more powerful research approaches; in many fields, computational approaches have expanded the scale of the questions that can be asked and increased the need for more and more diverse computational resources.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

UW-Madison researchers describe large, complex problems that require significant computational resources as current and emerging challenges for advancing discoveries in their fields. In the physical sciences, new models and approaches that merge simulations and experimental observations will be the future path to addressing big questions in areas such as plasma/fusion physics, material science, and economics. This is especially pronounced in fusion energy research where complex simulations that bridge experimental observations which may be incomplete or imperfect and simplified models have been developed, but using them to tackle the challenging questions requires significant computational resources. In this area, modeling done by UW-Madison physicists in support of ITER (a large international fusion reactor experiment, https://www.iter.org), FNSF (a fusion materials lifecycle testing facility), and DEMO (a prototype power producing fusion reactor) necessitates simultaneous modelling of materials, neutral particles, and ionized plasmas at high temperatures. While many codes accomplish individual parts of the puzzle, efforts are moving towards combining the existing codes together to give self consistent answers. The problems at hand are inherently multiscale: from neutron damage in the crystal lattice of tungsten on a pico-second scale up to device size (10s of meters) over the lifetime of a reactor. In addition, one of the most important challenges in plasma physics is quantitatively predicting the behavior of very complex magnetic confinement experiments, especially in future burning plasma fusion devices. The main efforts in this regard use fluid and/or particle simulations to model the plasma and atomic and materials simulations to model the plasma-wall interface and the device wall and blanket. For instance, the planned ITER experiment is expected to be vulnerable to violent plasma disruptions that in extreme cases could cause various kinds of damage to the device itself. Understanding before the fact how best to predict, avoid, control, and mitigate disruptions requires a comprehensive modeling effort, covering the entire system from plasma core to edge to the device wall, coupled with iterative experimental comparisons from present devices. This combination of simulations and experiments may allow quantitative validation of the physics models needed to describe ITER scenarios relevant to addressing disruptions, but the physics and computational challenges involved are formidable. The basic problem is that, because of factors like the large separation of scales important in plasma physics models, running full nonlinear simulations of the entire system is prohibitively expensive given present computational power. Thus, part of the research is in reduced models covering different parts of the system that are combined together to form a whole. However, even this simplified project is at the limit of present computational power, and in any case such modeling still needs experimental validation. In particular, in order to generate the required uncertainty estimates for the rigorous validation of nonlinear simulation models, it is necessary to do multiple code runs of scans for all of the input parameters, which, for systems with as many degrees of freedom as plasma models tend to have, very quickly becomes very expensive computationally.

While significant effort has been put into improvement of computational models used in fusion energy sciences, rigorous validation of these models is necessary in order to increase confidence in their ability to predict the performance of future devices. One UW-Madison researcher is presently using the NIMROD and DEBS codes on parallel computing systems to simulate plasmas to understand how magnetic fluctuations (which are related to plasma confinement) scale with the Lundquist number, a dimensionless quantity that represents the relative importance of Alfven wave dynamics to resistive dissipation. In the experiment, the Lundquist number can be orders of magnitude higher than can be simulated in a reasonable amount of time (spatial resolution and time steps must decrease as the Lundquist number is increased). To carry out the analysis, a set of simulations varying the Lundquist number across a lower range than accessible in the experiment are performed and uncertainty analysis must be carried out with an ensemble of runs with varied inputs.

In the life sciences, next generation sequencing is enabling analysis of genomes from numerous individuals at scales that will provide new insights into the interplay between an individual’s genetic makeup and the environment and/or physiological context. One UW-Madison botanist is sequencing hundreds of individuals in wild/natural populations (without reference genomes) to study population genetics while a microbial ecologist analyzes sequencing data for environmental samples to gain biological understanding from DNA fragments recovered from microbial communities ranging from soil to water to human guts and integrates this information with embedded environmental sensor datastreams. Another researcher is discovering genetic interactions that drive complex traits, such as obesity or diabetes. This work involves computing all pair-wise interactions in the genome and can quickly scale to billions of computations. The availability of increased computing power and resources will allow this approach to applied to many traits to more fully understand the implications of genetic interactions. In addition to the power next generation sequencing brings to addressing questions at this level, managing the large amount of data generated from sequencing microbial samples and also computing that data in a efficient manner are and will continue to be challenges.

In the medical sciences, computational methods are at a point of making many new clinical approaches possible. For example, the field of radiotherapy is now at a point where cutting-edge systems offer the ability to create new treatment plans for patients based on daily 3D images while the patient is in a treatment room and in a treatment position. New plans must be calculated quickly, and the users must be able to rely on a good plan being created on the first try. Computational research is required to ensure both that the computation is as fast as possible (whether analytic or Monte Carlo) and that optimization parameters reliably produce good treatment plans. Researching optimization parameters is a large endeavor, requiring thousands of treatment plan calculations per patient dataset to produce meaningful results. In addition, Implementing fast algorithms for clinical adoption of methods can be used to estimate tissue stiffness in vivo. This helps diagnose patients who are at high risk. For example a patient with soft tumor might have a cancer. In the field of material science, detailed numerical simulations that can capture the instantaneous physics of a multiphase flow will be important. Numerical methods for multi-physics aspects has still not reached a robust stage yet, and hence, it continues to be a challenge. Additionally, the analysis of the physics stemming from these simulations is scarcely being done. It remains more of an opportunity than a challenge. Another opportunity is determining the atomic structure of complex materials from incomplete, imperfect experimental data. One approach is to combine simulations of experimental data with simulations of the system energy into a global structural optimization scheme. In addition, discovery of interesting and useful signals in TB-scale (or larger) materials characterization data sets made possible by recent advances in instrumentation, will be an area of active research.

In other fields, UW-Madison researchers indicate that CryoEM imaging will become increasingly important in the study of structural biology and will be a big driver of large computations and Monte Carlo simulations of stellar and planetary dynamics and estimation of structural models of economic behavior, suitable for the analysis of the effects of alternative policies where no experimental data are available will be advanced by increased computational processing capabilities. Managing data also figures into the scientific challenges of several of these researchers, including integrating different types of data in a manner that respects domain knowledge and evaluating predictive models when there is a lack of ground truth and finding capabilities for managing data, given financial, staffing, and computing requirements.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The availability of resources for both high performance (HPC) and high throughput computing (HTC) is currently a limiting factor for UW-Madison researchers in a number of fields. High performance computational resources are essential to large scale nonlinear problems such as magnetized plasma physics because much of the important dynamics often happens on vastly smaller scales. Simulating effects across scales varying by more than 8 orders of magnitude requires an incredible number of computations and many real world problems have even greater scale separations. To have any hope of rigorously understanding these effects we need computers capable of solving these problems in a reasonable amount of time. A reasonable amount of time for simulations is generally thought of as a few days, since problems can have so much variability that simulations must be repeated many times over.
An experimentalist who makes use of advanced numerical physics simulation codes finds a basic need is computation speed for solving coupled nonlinear differential equations in large computational domains, with spatial resolutions reaching hundreds of thousands of points for millions of time steps. A useful nonlinear run with an extended physics model covering a few relaxation events might take several hundred thousand cpu hours on a modern parallel machine with a few hundred processors, and in wall time this might be several months. It is possible that the scale of present computational resources is sufficient along with a conglomeration of reduced models but also plausible that a set of full physics simulations could be run in a rigorous way on the scale of ten years of wall time, given serious advances in computational power.

A typical simulation with NIMROD to simulate magnetic fluctuations in plasmas requires 6 nodes each with 20 cores and generally take months of run time to finish. Typical resources needed are in the 100,000 to 500,000 CPU hours per run. Many runs are required to determine how quantities of interest change as input parameters (such as the Lundquist number) are varied. To truly understand uncertainty in quantities of interest, many more runs are necessary and to push simulations to higher Lundquist numbers, which are more relevant to experiments and future reactors, much more computational might is necessary. Determining the atomic structure of complex materials requires large-scale computing at the petaflop to exaflop scale, (but to make use of computing at that scale requires innovation in highly scalable algorithms and implementations for materials simulations), while mining large materials data sets requires large storage and high memory computing and the ability to move large data sets from local storage for acquisition to remote computing resources that are much less accessible than large scale computing right now.
The computing power and space needed to assemble and store the massive amounts of genetic data generated by next-generation sequencing are high. Resources like high-throughput computing clusters make de novo assembly possible, but wait times for jobs to run and file transfer times are long. Increased access to computing power is also the key infrastructure need to accomplish computing pair-wise interactions in genomics. Researchers currently split up the analysis and spread it over thousands of jobs but Increasing the size of the computing network would allow them to efficiently complete analyses more efficiently and apply these methods to more complex problems.

Another area that will drive future research in several domains is innovations in theory, algorithms, and computational strategies. For example, making use of large-scale computing at the petaflop to exaflop scale will depend on highly scalable algorithms for simulations in material science. Development of HPC tools for numerical partial differential equations (e.g., pressure-Poisson, hyperbolic systems for level set advection, and even parabolic system describing chemical and energy transport) would significantly benefit scientific and engineering efforts in general. As a necessary condition, these tools need to run efficiently in larger numbers of cores (~100,000 or greater), where issues related to parallelism and optimum memory and CPU usage are handled efficiently and behind the scenes. Currently, one of the major bottlenecks is that people with expertise in these HPC advancements are not the same people studying physics or developing models. Furthermore, the development of these tools will accelerate the advancement in all areas where the solution of these partial differential equations is a cornerstone, e.g., computational fluid dynamics. Additionally, new methods are needed for genome assembly from complex community datasets as well as new strategies for annotation and structure-function predictions and new theory for integrating sensor data with multi-'omics data. Large team projects conducted by teams using different computing systems, software environments, programming languages, etc. present a challenge, especially when trying to reconstruct an end-to-end data analysis pipeline in which different pieces may have been built and tested by different individuals. This problem is as much cultural as it is technical.

Several researchers note that access to accelerator resources are key to machine learning approaches with many parameters on large datasets and describe limitations with data storage, management, and movement. Finally, if the existing cyberinfrastructure were to be used clinically under a protocol, security issues (mainly secure data transfer) would need to be acceptable to meet institutional and regulatory guidelines. If code were developed in-house for clinical purposes, application/revision control and data fidelity would need to be addressed to ensure patient safety.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

UW-Madison researchers note the importance of investing in a diversity of resources and in professional staff who will steer researchers to the right resources, at both the campus and national levels. Computational diversity is important because even with the very powerful computers in existence today, that number is still dwarfed by how many problems there are to solve. Of course the most powerful systems can be highly expensive and therefore it is important to have a pool of computational resources that offer a range of performance abilities while also providing a complementary range of availability. This will allow scientists to focus on the most important aspects of their targeted problems and tailor their research to match the computational resources that are a best fit. Professional support staff are a key to finding creative solutions to problems by asking questions about what is the goal, where can computing help, what type of applications are already available, so researchers are not continually reinventing software to solve the same problems over and over. Very often faculty are familiar with but far from experts in the advanced computational environments available for research and facilitators at the UW-Madison Center for High Throughput Computing and other similar facilities at other institutions have been invaluable in troubleshooting and offering advice, and therefore ensuring work can move forward. Any initiative to develop or increase advanced computing resources should be done keeping in mind how important staffing and ancillary resources are to making large-scale institutional computing centers successful.

While a number of UW-Madison researchers utilize national computing resources, they point out the importance of on-campus resources. For example, one group has also been looking to use NERSC resources, but found accessing those computational resources is more difficult, requiring annual proposals and the readily available campus resources to be faster.Another group estimates that the computational power and queue times available on campus resources to be similar to those for the Edison machine at NERSC and at the same time, because of the greater access to run time on resources accessed on campus they are freer to experiment and iterate with different simulation conditions, which is very useful in their research context of validating simulation versus experiment.

Consent Statement